ChemGPS-NP: Tuned for Navigation in Biologically Relevant Chemical Space

Josefin Larsson,[†] Johan Gottfries,[‡] Sorel Muresan,[§] and Anders Backlund*[,†]

Division of Pharmacognosy, Department of Medicinal Chemistry, BMC, Uppsala University, Box 574, S-751 23 Uppsala, Sweden, Department of Medicinal Chemistry, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden, and GDECS Computational Chemistry, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden

Received January 3, 2007

Natural compounds are evolutionary selected and prevalidated by Nature, displaying a unique chemical diversity and a corresponding diversity of biological activities. These features make them highly interesting for studies of chemical biology, and in the pharmaceutical industry for development of new leads. Of utmost importance, for the discovery of new biologically active compounds, is the identification and charting of the corresponding biologically relevant chemical space. The primary key to this is the coverage of the natural products' chemical space. Here we introduce ChemGPS-NP, a new tool tuned for handling the chemical diversity encountered in natural products research, in contrast to previous tools focused on the much more restricted drug-like chemical space. The aim is to provide a framework for making compound classification and comparison more efficient and stringent, to identify volumes of chemical space related to particular biological activities, and to track changes in chemical properties due to, for example, evolutionary traits and modifications in biosynthesis. Physical−chemical properties not directly discernible from structural data can be discovered, making selection more efficient and increasing the probability of hit generation when screening natural compounds and analogues.

Traditionally, natural products have been essential sources of new drugs, and many drugs on the market were initially synthesized to mimic the actions of molecules found in Nature.[1,2] More than half of the new chemical entities introduced between 1981 and 2002 were natural products or natural product related, and the investigation of natural products as therapeutics in Western pharmaceutical industry reached its peak just before this period.[2]

During the 1990s, the leading source of new compounds in drug discovery shifted from natural products to high-throughput synthetic medicinal chemistry libraries. Combinatorial chemistry, through rapid assembly of various chemical units, provided large numbers of small, different but structurally related molecules for high-throughput screenings (HTS) against particular targets. The drug industry requested rapid screening and hit identification, while investigation of natural products involved time-consuming extract-library screening, bioassay-guided isolation, followed by laborious structure elucidation. HTS enabled screening of large defined libraries simultaneously at a rate that gave natural product research competitive disadvantages. The synthetic medicinal chemistry efforts together with HTS accelerated the synthesis, but unfortunately did not fulfill the expectations of increasing the number of lead candidates or drugs. Recently there seems to be a renewed interest in natural products in drug discovery, and the procedures used are becoming increasingly automated and made more effective.[3,4]

Natural products have a unique and vast chemical diversity and can be regarded as biologically explored, evolutionarily selected, and prevalidated by Nature. The probability that natural products and derived structures will be biologically relevant is subsequently high. Virtually all of the biosynthesized compounds have some beneficial purpose for the producing organism, e.g., receptor binding capacity,[2,5] in order to interact with multiple proteins and thereby eventually compete for resources, e.g., by avoiding predation. Natural products also astonishingly often have advantageous pharmacokinetic properties as a result of their mission of identifying targets in the organism.[6] These facts support the hypothesis that natural products are exceptional design resources in drug discovery and that screening natural products should yield high hit-rates. Of utmost importance for discovery of new active compounds for future therapies is the identification and charting of a biologically relevant chemical space, and a primary key to this is the coverage of the diverse natural products' chemical space.[2,7−9]

The concept, chemical space, is often used instead of "multidimensional descriptor space", which is a region defined by the descriptors chosen to describe a set of chemicals.[10] Chemical space can be compared to cosmic space, where chemical compounds, instead of stars, occupy the space. It is enormous and basically infinite, comprising all possible molecules. The total number, only of small carbon-based compounds, is estimated to exceed 10^{60},[11] a number that will rise significantly when adding larger and more complex bioactive molecules. A typical library file for the top pharmaceutical companies contains at most a few million compounds,[12] which, following this discussion, offers only a modest sampling of all the potential compounds that comprise chemical space and most presumably comprise historic bias regarding diversity. Considering the enormous number of possible compounds and the given capabilities of HTS, the process of compound selection and prioritization is crucial.[13] To define the biologically active chemical space, the entire chemical space first needs to be populated to some extent.[10] There is a critical need for tools able to chart biologically relevant chemical space and provide an efficient mapping device for selection of high-probability hits and prediction of their properties and activities. ChemGPS-NP[9] is a tool fulfilling these requirements. It is tuned for exploration of the regions of chemical space most likely to enclose compounds with activities of interest, the biologically relevant natural products' (NP) chemical space. ChemGPS-NP also has the capacity of serving as a reference system enabling characterization and comparison of molecules from various research groups.

A chemical global positioning system (ChemGPS) based on medicinal chemistry drug-like molecules[14] has in a previous study,[9] when applied to a set of natural products with cyclooxygenase inhibiting activity, encountered numerous outliers, i.e., extrapolations predicted outside the model. This result could indeed be expected since natural products are extremely diverse and some-

* To whom correspondence should be addressed. Tel: +46-18-4714498. Fax: +46-18-509101. E-mail: anders.backlund@fkog.uu.se.
[†] Uppsala University.
[‡] Department of Medicinal Chemistry, AstraZeneca R&D Mölndal.
[§] GDECS Computational Chemistry, AstraZeneca R&D Mölndal.

10.1021/np070002y CCC: $37.00
Published on Web 04/18/2007

times very different in terms of structure and chemical properties as compared to the much more restricted drug-like chemical space for which ChemGPS was designed. Predicted outliers result in extrapolation and thus uncertainty in accuracy and precision, which is avoided using ChemGPS-NP.

A map of chemical space can be constructed by applying the same principles as the Mercator convention in geography: Rules correspond to dimensions (e.g., longitude and latitude), and structures correspond to objects (e.g., cities and countries).[14,15] Objects include a set of satellite structures similar to the satellites used in the Navstar global positioning system[16] and a set of core structures. Satellites are intentionally placed outside the natural products' chemical space by having extreme values for one or several of the desired properties and thereby marking the limits of the chemical space of interest. In this way the positions that other compounds may have are well covered. With ChemGPS-NP we attempted to better represent the entire biologically relevant chemical space. This was achieved by including complex structural examples from the creative chemistry of Nature's bioactive molecules. In ChemGPS-NP selected main rules include aspects of size, shape, lipophilicity, polarity, polarizability, flexibility, rigidity, and hydrogen bond capacity. The rules and objects together present a chemical space map. The ChemGPS-NP space map coordinates are *t*-scores from principal component analysis (PCA)[17,18] using a carefully selected subset of 35 descriptors that evaluate the above-mentioned rules on a total set of 1779 chosen satellite and core structures. Using ChemGPS-NP, novel structures are positioned in chemical space via PCA score prediction. This overcomes component rotation drawbacks of local models, which in addition needs to be recalculated whenever a new set of compounds are added or removed. Furthermore, ChemGPS-NP is a global model and thus amenable to comparison with other models. Hence, it may serve as a reference system by which large libraries can be compared without scores changing as new structures are included and predicted. It handles novel compounds via interpolation and avoids extrapolations since the principal property space is well covered, in all directions, by relevant satellite structures.

The natural products occupy a different and larger space than that normally dealt with in medicinal chemistry.[7,9] Feher and Schmidt[7] compared representative combinatorial, synthetic, and natural product compound libraries on the basis of molecular diversity and "drug-likeness" properties. Other groups have distinguished natural products, drugs, or other synthetic compounds on the basis of structural similarity,[8,19] pharmacophore properties,[19] or other molecular descriptors.[20] These studies show that natural products typically have a greater number of chiral centers and increased molecular complexity as compared to synthetic drugs and combinatorial libraries.[7] Furthermore they often contain fewer nitrogen, halogen, and sulfur atoms, but are noticeably more oxygen-rich.[7,20] Natural products also differ by having a higher number of hydrogen bond donors and acceptors, by containing a larger number of rings, and by being more structurally rigid. Additionally, they have a broader distribution of, for example, molecular mass, octanol−water partition coefficient, and diversity of ring systems compared to synthetic and medicinal chemistry compounds.[7,19,20]

Results and Discussion

Dictionary of Natural Products (DNP), released October 2004 comprising 167 169 compounds, was used as a starting set. In a first step, the compounds (represented as SMILES) were pruned of duplicate or erroneous data, resulting in 124 082 unique structures. Subsequently 712 compounds with elements other than H, C, N, O, F, P, S, Cl, Br, or I were removed, giving a set of 123 370 substances. Cluster analysis of the remaining DNP compounds was performed, resulting in 10 859 clusters, of which 2307 were singletons, and 1376 clusters contained more than 50 substances. The cluster seeds of these 1376 clusters were selected as a starting model for the emergent ChemGPS-NP.

As only 1D and 2D descriptors can be calculated from SMILES, it is not possible to determine absolute configuration, and as a result this does not form part of the overall calculation. Other studies have demonstrated that such an approach is valid, in that those analyses employing 3D molecular descriptors generally do not perform any better than those using 2D descriptors.[21−23] Also, for many natural products the absolute and relative configuration have not been determined.

A set of 926 molecular descriptors was primarily calculated for all compounds used in this study, after which the molecular descriptor array was pruned on the basis of a number of criteria. For inclusion a descriptor should (1) have a comprehensible physical meaning (descriptors that were not intuitively easy to interpret were removed, since the goal was to deliver a tool that facilitates understanding and explanation of chemical space); (2) reveal loading in at least one component of the principal component analysis (PCA)[17,18] model when terminated with statistical cross-validation criterion (see below); (3) be able to distinguish between the compounds (a descriptor whose value varies little within compounds in a data set has little power to make a distinction between these compounds[20]); (4) encode relevant aspects of molecular complexity; and (5) describe at least any of the following intuitively important molecular properties: lipophilicity, polarity, size/shape, hydrogen bond capacity, polarizability, flexibility, and rigidity.

The final set of 35 descriptors[24] is presented in Table 1. Lipophilicity was estimated by the Ghose−Crippen ALogP.[25−27] Polarity was estimated using descriptors calculating topological polar surface area using nitrogen and oxygen contribution or nitrogen, oxygen, phosphor, and sulfur contribution [TPSA (NO) and TPSA (Tot)],[28] hydrophilic factor (Hy),[29] and the counts for oxygen (nO), aliphatic/aromatic hydroxyl groups (nROH/nArOH), and nitrogen (nN). Size- and shape-related descriptors included molecular weight (MW), number of atoms (nAT), number of carbons (nC), number of non-H atoms (nSK), and the Ghose−Crippen molar refractivity (AMR).[25] Hydrogen bond capacity was measured by counting the number of nitrogen and oxygen as donor atoms for hydrogen bonds (nHDon) and the number of nitrogen, oxygen, and fluorine as acceptor atoms for hydrogen bonds excluding nitrogen with positive formal charge, higher oxidation states, and the pyrrolyl form of nitrogen (nHAcc). In addition, nO, nN, and nROH/nArOH were used to estimate this capacity. Polarizability was taken into account through summing atomic polarizabilities (Sp) and calculating AMR. Flexibility and rigidity were estimated by counting the total number of bonds (nBT), rings (nCIC), and rotatable bonds (RBN)[30] and by calculating the rotatable bond fraction (RBF). The constitutional descriptors—sum of atomic van der Waals volumes (Sv), sum of atomic Sanderson electronegativity (Se), mean atomic van der Waals volume (Mv), mean atomic Sanderson electronegativity (Me), number of non-H bonds (nBO), number of multiple bonds (double, triple, and aromatic bonds) in a molecule (nBM), aromatic ratio (ARR), number of double bonds (nDB), number of aromatic bonds (nAB), number of halogens (nX), and number of benzene-like rings (nBnz)—were added together with the functional group counts: number of aromatic carbons (sp^2) (nCar) and number of amides (n_amid). As a tool for discussions, the often referenced molecular property descriptor known as Lipinski alert index (LAI)[31,32] was included.

PCA[17,18] was used to analyze the multidimensional data and to create a global model of biologically relevant chemical space. PCA can be used to find patterns in large data sets by filtering out noise and reducing the dimensionality. Correlated variables are compressed into a smaller number of new uncorrelated variables, principal components (PCs), while retaining as much information as possible. PCA is expressed in terms of scores and loadings, where the scores are related to the objects and the loadings are related to

Table 1. ChemGPS-NP Descriptors

number	abbreviation	description
10	MW	molecular weight
2	Sv	sum of atomic van der Waals volumes (scaled on C atom)
3	Se	sum of atomic Sanderson electronegativities (scaled on C atom)
4	Sp	sum of atomic polarizabilites (scaled on C atom)
5	Mv	mean atomic van der Waals volume (scaled on C atom)
6	Me	mean atomic Sanderson electronegativity (scaled on C atom)
7	nAT	number of atoms
8	nSK	number of non-hydrogen atoms
9	nBT	number of bonds
10	nBO	number of non-hydrogen bonds
11	nBM	number of multiple bonds
12	ARR	aromatic ratio
13	nCIC	number of rings
14	RBN	number of rotatable bonds
15	RBF	rotatable bond fraction
16	nDB	number of double bonds
17	nAB	number of aromatic bonds
18	nC	number of carbon atoms
19	nN	number of nitrogen atoms
20	nO	number of oxygen atoms
21	nX	number of halogens
22	nBnz	number of benzene-like rings
23	nCar	number of aromatic carbon atoms (sp^2)
24	n_amid	number of amides
25	nROH	number of aliphatic hydroxy groups
26	nArOH	number of aromatic hydroxy groups
27	nHDon	number of donor atoms for hydrogen bonds (N and O)
28	nHAcc	number of acceptor atoms for hydrogen bonds (N, O, and F)
29	Ui	unsaturation index
30	Hy	hydrophilic factor
31	AMR	Ghose−Crippen molar refractivity
32	TPSA(NO)	topological polar surface area using N and O
33	TPSA(Tot)	topological polar surface area using N, O , S, and P
34	ALOGP	Ghose−Crippen octanol−water partition coefficient
35	LAI	Lipinski alert index (drug-like index)

the variables. The results can be viewed in score and loading plots, where the relative distance between compounds in chemical space becomes a measure of their similarity with respect to the particular set of descriptors considered. The optimal number of PCs in the ChemGPS-NP model was decided using cross-validation, with the additional criterion that any descriptor should load in at least one PC. Prior to PCA all data were centered and scaled to unit variance.

For the present study, 46 diverse compound sets comprising more than 1 million unique compounds were compiled for the subsequent validation (Supporting Information, Table 1). Special efforts were made to include chemical substances of diverse biological origin, including bacteria and eukaryotes from several phyla, as well as organisms from different ecological niches.

Distance to the model in the X space[33] after selecting the number of components for new observations in the prediction set, DModXPS, was calculated for all compounds predicted with the evolving ChemGPS-NP and expressed as normalized distances in units of standard deviations (SDs). To progressively expand the ChemGPS-NP chemical space, each of the 46 data sets were predicted with the current model as a training set. Compounds in a prediction set with a predicted DModX larger than four SDs were considered outliers, i.e., significantly different from the compounds used to construct the model. Every such set of outliers would potentially enhance the ChemGPS-NP coverage of the drug space and successively yield a convergent ChemGPS-NP. The identified outliers were all individually scrutinized. If there were less than 20 interesting outliers in a data set, they were all added to the training set. If the prediction set contained more than 20 interesting outliers, a subset was selected via the D-optimal design,[34] or, if they were more than 100, via D-optimal onion design.[35,36] D-optimal designs select the most extreme points of the candidate set and give a minimal set of selected compounds with maximum diversity. D-optimal onion designs divide the set into a number of selected layers where one separate D-optimal design is made in each layer. Thereby it samples more evenly throughout the region. Selected outliers were included in the next version of the training set, ChemGPS-NP. This procedure was iterated until all 46 data sets had been predicted.

Starting from the first training set of ChemGPS-NP derived from DNP as explained above, the 35 ChemGPS-NP descriptors were calculated. PCA was performed on the resulting data matrix. As a first expansion of ChemGPS-NP a larger portion of the medicinal chemistry training set ChemGPS[14] was included to cover this overlapping space. As the training set should be as diverse as possible, we attempted to choose molecules from ChemGPS that were as different as possible from the first version of ChemGPS-NP. ChemGPS was predicted with ChemGPS-NP, and 283 compounds that had a DModXPS larger than the critical value at a probability level of 5% (here 1.17 SDs) were included in a new version of ChemGPS-NP.

The ChemGPS-NP descriptors were subsequently calculated for the next of the 46 data sets, and this set was positioned with PCA score prediction based on the scores of the training set, ChemGPS-NP. DModXPS was calculated, and the results were listed and sorted in order of decreasing value. All compounds with a DModXPS higher than four SDs were inspected and eventually included in the training set as explained above. ChemGPS-NP was then recalculated, and a new model was made with the selected outliers included in the training set. A new data set was selected as prediction set, and the same procedure was iterated. In the beginning many compounds were added each round, and new models calculated, but after a number of iterations there were no additional outliers encountered according to the criterion.

The operational ChemGPS-NP includes 1779 compounds and has predicted 619 382 compounds, without encountering any outliers. The PCA modeling was terminated at eight PCs, after inspection of cross-validation and loading vectors, as described above (R^2X was 0.92 and Q^2(cum) was 0.73). A summary of the eight PCs of the model is shown in Figure 1.

The interpretation of the dimensionality includes the weight of individual variables in the data set, which indicate what principal molecular properties are explained by the respective orthogonal components (i.e., the PCs). The four most significant PCs explain 77% of the variance and can be interpreted as follows: PC1 represents size, shape, and polarizability, PC2 corresponds to aromatic- and conjugation-related properties, PC3 describes lipophilicity, polarity, and H-bond capacity, and PC4 expresses flexibility and rigidity. For a thorough description of PC5−PC8 we refer to Supporting Information Figure S1. Figure 2 shows the score plot of the three most significant PCs with some of the extreme chemical objects encircled and illustrated in Figure 3.

The ChemGPS-NP model is not limited to the present choice of descriptors. The ChemGPS-NP descriptors described in this paper were successfully replaced with VolSurf descriptors[37] in order to obtain consistent maps of the drug-like chemical space (data not shown). VolSurf descriptors are based on 3D representations of the included molecules and their surface properties, which constitute an alternative and different starting point for the molecular description. In this validation process VolSurf descriptors were calculated for ChemGPS-NP. Latent variables were extracted by applying PCA to the VolSurf descriptors set, and the main rules that could be calculated with VolSurf were compared with the

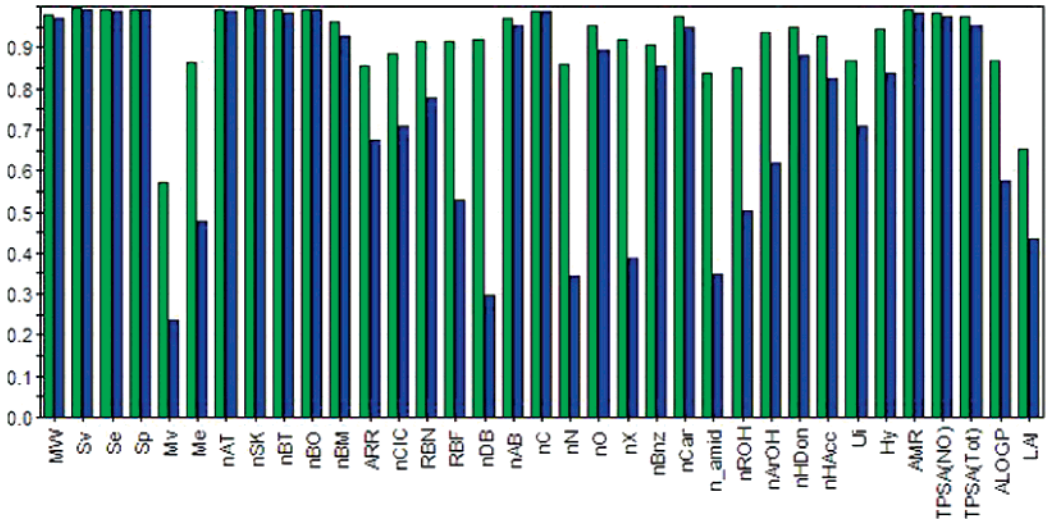

Figure 1. Summary of the eight components of ChemGPS-NP. The *x*-axis gives the variable IDs, i.e., the abbreviations for the descriptors explained in Table 1. The *y*-axis denotes the cumulative explained variance by regression (R^2, green) and by cross-validation (Q^2, blue) for the descriptor matrix including the 35 variables (molecular descriptors) after 8 principal components.

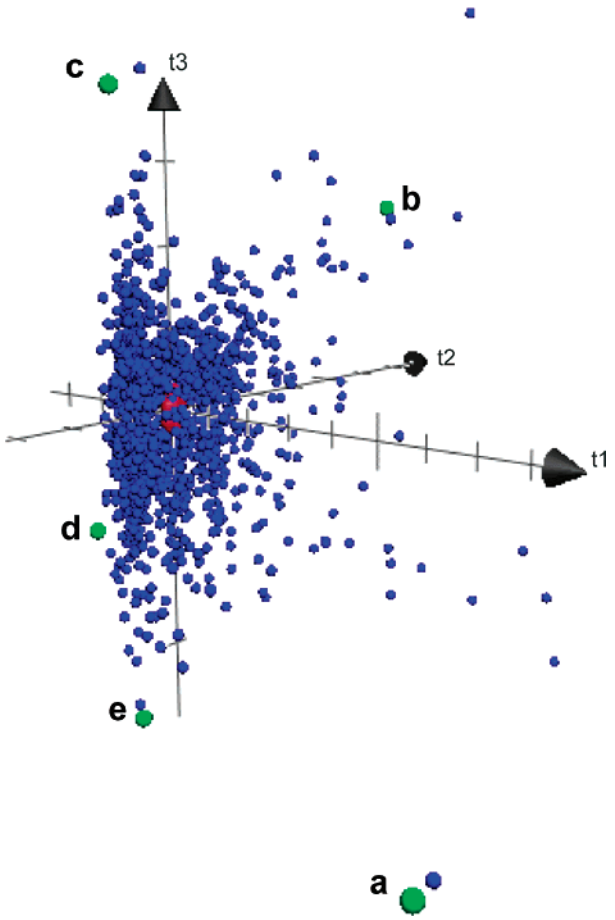

Figure 2. Score plot of the three most significant dimensions (t1, t2, and t3) of ChemGPS-NP, illustrating the general shape of natural products chemical space, revealing its prominent parametrical asymmetry. Each sphere represents an object (a compound). Green spheres (a−e) are illustrated in Figure 3. The first three PCs explain 71% of the variance.

DRAGON-related scores. This comparison, using 2D- and 3D-based description, respectively, indicated that similar molecular property dimensions were found by both approaches, which in turn validated that a robust molecular principal property space has been established.[14]

Previous work has addressed a ChemGPS aimed for medicinal compounds,[14] i.e., compounds within or near the so-called Lipinski space.[31] The ChemGPS led to an estimation of a chemical space of six to eight dimensions, as determined by PCA. The inherent engine of PCA comprises a search for data variance in new orthogonal dimensions, which are ranked by their quantitative level of explained variance (i.e., as estimated by R^2 and Q^2), such that the first principal component explains more variance than the second, which in turn explains more than the third, etc. Interestingly, in ChemGPS, size and shape were explained in PC1, lipophilicity-related parameters were described in PC2, and flexibility versus rigidity and polar variables were explained in PC3.[14] In the present study an expansion was made by shifting the focus toward natural products, and thereby a different order of explained properties was revealed. In ChemGPS-NP lipophilicity is not explained until in PC3, replaced as PC2 by aromaticity- and conjugation-related properties of the compounds, while flexibility and rigidity properties are explained in PC4. This is the most prominent manifestation of differences between drug-like and natural products chemical space. There can be several possible explanations for such a switch, including the fact that medicinal chemists more often tend to explore hydrophobic interactions between ligands and biological targets.[31] This would subsequently lead to an artifactual increase in variation in lipophilicity, resulting in a comparably larger variation in this respect dealing with man-designed molecules. On the other hand, one can see it from Nature's perspective where evolutionary pressure is the major driving force. Natural compounds in general and secondary metabolites in particular are bound to function in a generally hydrophilic environment. In order to retain, for example, supposed defense substances in solution, highly lipophilic substances must be avoided and, hence, the variation in lipophilicity is reduced by a functional constraint. With a lower degree of variation follows a lower explained variance and consequently a lower order of the corresponding principal component.

Considering dimensionality, the ChemGPS-NP model requires eight rather than six PCs in order to reach an acceptable level of explanation. The reason for the higher dimensionality of natural products as compared to drug-like compounds appears logical. The interactions/models employed in the drug industry are, for good reasons, usually designed to address a simple and clear-cut situation, a constraint seldomly encountered in Nature. More complexity results in higher dimensionality. The reason eight dimensions are sufficient for the ChemGPS-NP is a tradeoff between acceptable explanatory power and comprehendible complexity.

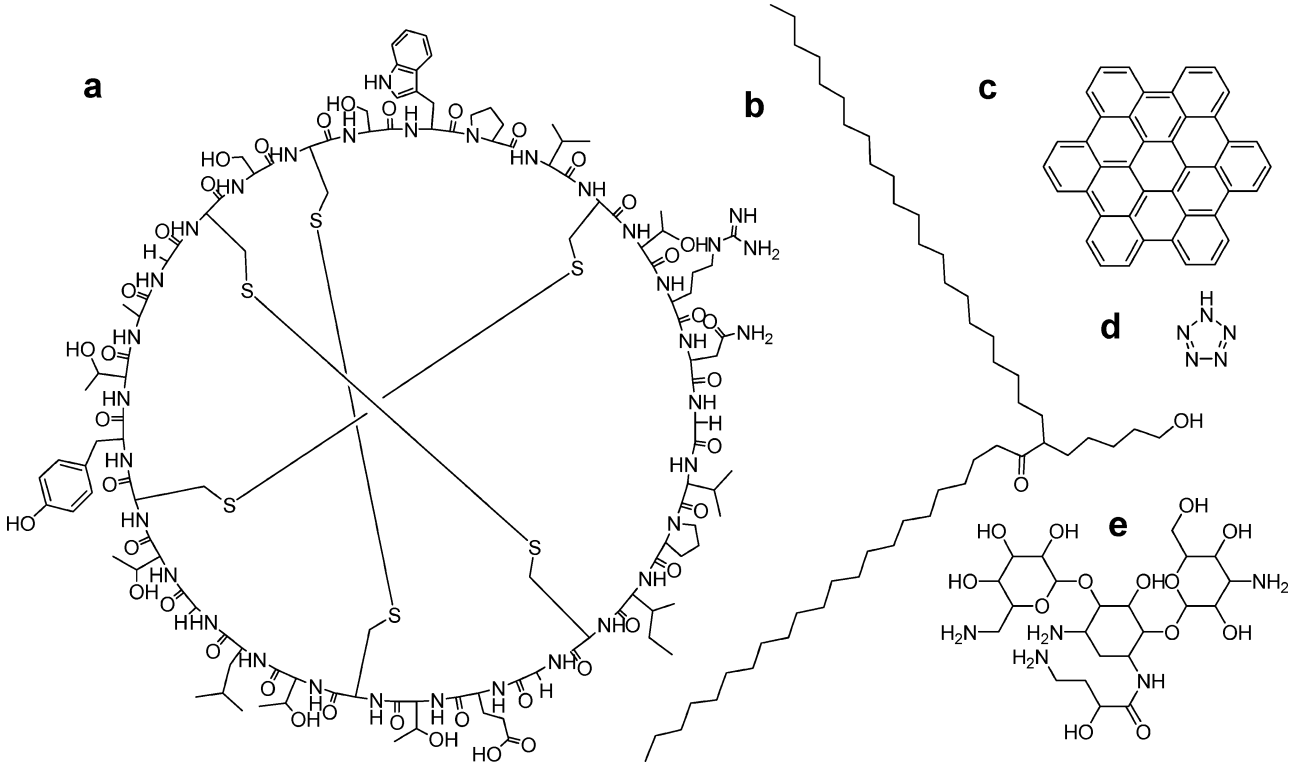

Figure 3. Molecular structures and ChemGPS-NP coordinates for the selected objects in Figure 2 (a−e). (a) varv F: t1 = 46.4, t2 = −10.2, t3 = −5.84. (b) 23-(5-Hydroxypentyl)-22-pentatetracontanone: t1 = 4.83, t2 = −4.52, t3 = 7.58. (c) Hexabenzocoronene: t1 = 4.33, t2 = 12.0, t3 = 4.79. (d) Pentazole: t1 = −4.82, t2 = −2.32, t3 = −2.98. (e) Amikacin: t1 = 5.01, t2 = −3.85, t3 = −6.89.

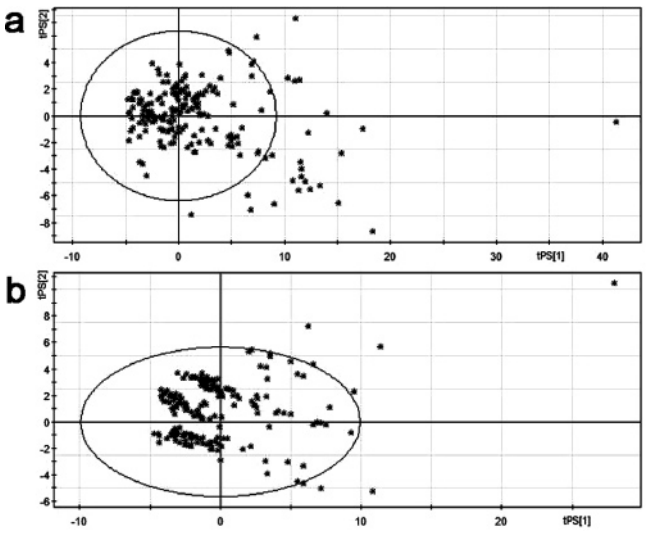

Figure 4. Comparison of the first two dimensions of the predicted score plots (tPS1 and tPS2) for collection of natural cyclooxygenase inhibitors[9] predicted using (a) ChemGPS and (b) ChemGPS-NP. More distinct clustering and fewer outliers are obtained with ChemGPS-NP.

Reanalysis of the previously mentioned set of natural cyclooxygenase inhibitors[9] with ChemGPS-NP indicates several interesting differences in interpretation, as compared to ChemGPS. ChemGPS-NP is more tuned to catch variations among natural products, and primarily, the data organize in more distinct clusters. Our conclusions from the previous study agree with what we can interpret from the present study, but as illustrated in Figure 4, we are now able to draw more detailed and informative conclusions from more well-defined clusters, without having to deal with outliers and extrapolation. This further emphasizes the need for ChemGPS-NP in natural product related research.

The drug discovery process is today hampered by increasing costs and high attrition rates, with an overall decrease in the number of annually registered new chemical entities. One way to try to overcome these obstacles would be a more efficient selection process, giving a higher probability of obtaining a lead compound. As discussed above, the use of evolutionary processes in Nature as a billion year prescreen is one possibility, but to further improve this, a method for navigation in chemical space would be necessary. The benefits of ChemGPS-NP are, in one way, comparable to the possibilities opened in molecular biology by rigorous application of the BLAST algorithms.[38] These allow, for example, through Web interfaces, the research community to easily compare sections of nucleotide or amino acid sequences for homology searching, identifying genes, or preparing data sets for phylogenetic analyses, all in huge data sets. It provides compound property description, clustering overview and property interpretation via the PCA loading vectors, directly amenable to deriving a global similarity metric. ChemGPS-NP is a tool tuned for managing the chemical diversity addressed in natural products research. It forms a framework for making compound comparison and selection more efficient and stringent. Physical−chemical properties not directly discernible from structural data can be discovered and quantified. Different volumes of chemical space corresponding to specific biological activities can be identified, which can be used for prediction or validation of both single structures and large compound sets. ChemGPS-NP also gives possibilities to interpret evolutionary driven changes in series of chemical compounds, effectively tracking the evolution of physical−chemical properties, and not only in modifications of compound structures. All of these features increase the probability of hit generation when screening the vast diversity of natural products in the search for novel bioactive molecules. ChemGPS-NP is still a heuristic model, and with future use it appears unavoidable that occasional outliers will show up. However, applying statistical molecular design procedures has maximized the probability of a coherent version and, thus, minimized the probability of novel satellites showing up in the future. Furthermore, by systematically predicting numerous relevant structures from well-

known sources (i.e., more than one million compounds of biologically diverse origins) by the ChemGPS-NP, which in a reproducible manner could predict their property positioning in chemical space without extrapolation, points toward a robust prediction engine. It is obvious already from work in progress to implement the model that ChemGPS-NP has potential to help gain novel chemical and biological insight in numerous ways.

Experimental Section

All molecular editing and filtering steps were performed with tools based on Daylight toolkit.[39] Cluster analysis was performed using Daylight fingerprints and a Tanimoto coefficient of 0.7 as similarity cutoff. Molecular descriptors were calculated from SMILES 2D representations for all compounds used in this study employing the software Dragon Professional, version 5.3.[40] All multivariate models were obtained using PCA as implemented in SIMCA-P+ 10.5.[33] DModXPS was calculated for all compounds predicted with the evolving ChemGPS-NP using SIMCA-P+ 10.5.[33] D-optimal (onion) designs[36] were generated with the software MODDE 7.[41] Design factors were scaled to unit variance and centered by default prior to the design.

Acknowledgment. This work was supported, in part, by the Swedish Research for Environment, Agricultural Science and Spatial Planning, grant number 229-2004-1714.

Supporting Information Available: Table listing the data sets used to develop ChemGPS-NP v.1 and figure illustrating the interpretation of PC5−PC8 are available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Cragg, G. M.; Newman, D. J.; Snader, K. M. *J. Nat. Prod.* **1997**, *60*, 52−60.
(2) Newman, D. J.; Cragg, G. M.; Snader, K. M. *J. Nat. Prod.* **2003**, *66*, 1022−1037.
(3) Koehn, F. E.; Carter, G. T. *Nat. Rev. Drug Discovery* **2005**, *4*, 206−220.
(4) Ortholand, J. Y.; Ganesan, A. *Curr. Opin. Chem. Biol.* **2004**, *8*, 271−280.
(5) Clardy, J.; Walsh, C. *Nature* **2004**, *432*, 829−837.
(6) Breinbauer, R.; Vetter, I. R.; Waldmann, H. *Angew. Chem., Int. Ed.* **2002**, *41*, 2879−2890.
(7) Feher, M.; Schmidt, J. M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.
(8) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. *Proc. Natl Acad. Sci. U.S.A.* **2005**, *102*, 17272−17277.
(9) Larsson, J.; Gottfries, J.; Bohlin, L.; Backlund, A. *J. Nat. Prod.* **2005**, *68*, 985−991.
(10) Dobson, C. M. *Nature* **2004**, *432*, 824−828.
(11) Bohacek, R. S.; McMartin, C.; Guida, W. C. *Med. Res. Rev.* **1996**, *16*, 3−50.
(12) Lipinski, C.; Hopkins, A. *Nature* **2004**, *432*, 855−861.
(13) Linusson, A.; Gottfries, J.; Lindgren, F.; Wold, S. *J. Med. Chem.* **2000**, *43*, 1320−1328.
(14) Oprea, T. I.; Gottfries, J. *J. Comb. Chem.* **2001**, *3*, 157−166.
(15) Oprea, T. I. *Curr. Opin. Chem. Biol.* **2002**, *6*, 384−389.
(16) Pike, J. 2006. Information about Navstar GPS available at: http://www.fas.org/spp/military/program/nav/gps.htm.
(17) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. PCA. In *Multi- and Megavariate Data Analysis*; Umetrics Academy: Umeå, 2001; pp 43−69.
(18) Pearson, K. *Philos. Mag.* **1901**, *2*, 559−572.
(19) Lee, M. L.; Schneider, G. *J. Comb. Chem.* **2001**, *3*, 284−289.
(20) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.
(21) Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Model.* **1996**, *36*, 572−584.
(22) Matter, H.; Potter, T. *J. Chem. Inf. Model.* **1999**, *39*, 1211−1225.
(23) Sheridan, R. P.; Kearsley, S. K. *Drug Discovery Today* **2002**, *7*, 903−911.
(24) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinham, 2000; Vol. 11, p 667.
(25) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(26) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(27) Viswanadhan, V. N.; Reddy, M. R.; Bacquet, R. J.; Erion, M. D. *J. Comput. Chem.* **1993**, *14*, 1019−1026.
(28) Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(29) Todeschini, R.; Vighi, M.; Finizio, A.; Gramatica, P. *SAR QSAR Environ. Res.* **1997**, *7*, 173−193.
(30) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. *J. Med. Chem.* **2002**, *45*, 2615−2623.
(31) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. *J. Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(32) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. *J. Adv. Drug. Delivery Rev.* **2001**, *46*, 3−26.
(33) *SIMCA-P+ 10.5*; Umetrics AB: Umeå, Sweden. Information available at http://www.umetrics.com.
(34) de Aguiar, P. F.; Bourguignon, B.; Khots, M. S.; Massart, D. L.; Phan-Than-Luu, R. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 199−210.
(35) Olsson, I.-M.; Gottfries, J.; Wold, S. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 37−46.
(36) Olsson, I.-M.; Gottfries, J.; Wold, S. *J. Chemom.* **2004**, *18*, 548−557.
(37) Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa, B. *THEOCHEM* **2000**, *503*, 17−30.
(38) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403−10.
(39) *Daylight 4.9*, Daylight Chemical Information Systems, Inc.: Santa Fe, NM. Information available at http://www.daylight.com.
(40) *Dragon Professional 5.3*; Talete srl: Milano, Italy. Information available at http://www.talete.mi.it/dragon.htm.
(41) *MODDE 7*; Umetrics AB: Umeå, Sweden. Information available at http://www.umetrics.com.

NP070002Y